DOCUMENT RESUME

ED 110 468 TM 004 739

AUTHOR Hill, Richard K.
TITLE Minimizing Context Effect When Using Multiple Matrix Sampling.
PUB DATE Apr 75
NOTE 10p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Washington, D.C., March 31-April 2, 1975)

EDRS PRICE MF-$0.76 HC-$1.58 PLUS POSTAGE
DESCRIPTORS *Bias; *Item Sampling; Matrices; Standardized Tests; *Statistical Analysis; *Testing; Testing Problems
IDENTIFIERS *Multiple Matrix Sampling

ABSTRACT

This study is an a priori demonstration of the applicability of multiple matrix sampling techniques to the practical research problem of parameter estimation. Three tests were administered to two separate but parallel populations, with one receiving item samples and the other receiving full tests. Special efforts were made to minimize the context effect due to sampling procedures. Parameters estimated from matrix sampling statistics closely matched those estimated from full test results, indicating little context effect bias due to the sampling procedures. (Author)

MINIMIZING CONTEXT EFFECT WHEN USING MULTIPLE MATRIX SAMPLING

Richard K. Hill

California State Department of Education

Objectives of the Inquiry. Multiple matrix sampling is rapidly gaining popularity as a tool for evaluation and research, but to date the development of the theory has far outstripped the practical application of it. A review of the multiple matrix sampling literature shows that a small minority of the studies conducted to date have been performed a priori. Most have been concerned with verification of the theory for estimating parameters from matrix samples, and thus have been either post hoc or Monte Carlo studies.

Although several studies have demonstrated that, in theory, parameter estimation can be accomplished through matrix sampling, little evidence exists that the approach works in practice. There is a need to develop and verify procedures to be used with multiple matrix sampling which will minimize the practical problem of context effect. The purpose of this paper is to propose some guidelines for use when applying multiple matrix sampling techniques, and to empirically test the effectiveness of these guidelines with a variety of test types.

Method. In this experiment, three tests were administered to fourth and fifth grade pupils in selected schools throughout New York and Pennsylvania. The tests were:

1. A 25-item short form of the Otis-Lennon Mental Ability Test, Intermediate II level.

The author wishes to recognize the cooperation and assistance of the staff of the Eastern Regional Institute for Education in conducting the study.

Paper presented at the National Council on Measurement in Education Annual Meeting, Washington, D. C., 1975.

2. The 60-item Intermediate I Science subtest of the Stanford Achievement Test battery.

3. A 35-item test, the Eastern Regional Institute for Education (ERIE) Science Process Test, designed by ERIE personnel to measure objectives of the Science--A Process Approach (SAPA) curriculum.

There were several reasons for selecting these tests. The Otis-Lennon and the Science Achievement tests are widely used standardized tests. Previous testing by ERIE showed them to be of approximately 50 per cent difficulty level for fourth graders, and since fifth graders were to be used in the study also, the distribution of scores was expected to be negatively skewed. The ERIE Science Process test had been very difficult for fourth graders in previous testings (results from the previous year had yielded a mean of 12.4 for the 35 items) and the distribution of test scores was expected to be severely positively skewed.

In virtually all matrix sampling studies to date, a major obstacle to clear interpretation has been that subjects received both the item samples and the total test. But in this study, no subject who received an item sample received the total test, and vice versa. Two separate but parallel populations were generated by drawing two fourth and two fifth grade classes in each of 14 schools. A class in each grade from each school was administered the item samples, with the other two classes receiving the full tests. Thus, no examinee was contaminated by being tested under both conditions, and yet parameters estimated through matrix sampling procedures could be compared to those estimated by full test administration.

The administration of the tests was designed to minimize the other major difficulty encountered in many matrix sampling studies--violation of the assumption of no context effect, which is composed largely of two factors--speededness and fatigue. In order to minimize fatigue effects, the students who were given the full tests received them over a two day period; the Science Process Test was administered the first morning, the Otis-Lennon Aptitude Test that afternoon, and the Science Achievement Test the following morning. The times allowed to complete the tests were 45, 30 and 45 minutes respectively; which were considered ample. The time allowed for students taking the item samples was 20 minutes, which was considered to be a generous estimate also. Thus, effects of speededness should have been minimal.

The item samples were chosen so that all samples were mutually exclusive and exhaustive. Five samples of five items each were chosen from the Otis-Lennon; seven samples of five items each from the Science Process Test; and ten samples of six items each from the Science Achievement Test.

The item samples from the tests were combined so that each possible combination of the samples from the three tests occurred at least twice, but no more than three times (there were a total of $[5x7x10] = 350$ possible combinations). Booklets then were constructed with the items from the Otis-Lennon as item numbers 1-5, the Science Process Test 6-10, and the Science Achievement 11-16. Each student knew only that he was receiving a 16-item test; no mention was made of the fact that his item sample was composed of three different tests.

Data Sources. After pretesting to assure that the directions were clear and the time limits were reasonable, the tests were administered to fourth and fifth grade students in the selected schools. The answer sheets were returned to the author for scoring and analysis.

Results. A total of 602 pupils took the full battery of tests, while 653 took the item samples. For each of the three tests, the mean and standard deviation were calculated. For the item samples, the statistics necessary to estimate total test mean and standard deviation were calculated. These results are displayed in Tables 1-3.

In each case, total test variance was estimated two different ways. The first estimate was calculated using an equation credited to Lord (1960):

$$V(X) = M \cdot V(X_i)\left[1 + (M-1)KR20\right], \qquad (1)$$

Where $V(X)$ = the estimated total test variance
M = the number of item samples
$V(X_i)$ = the variance of the ith item sample
KR20 = the value of KR20 for the ith item sample

The second estimate was calculated using an equation from Hill (1972):

$$V(X) = M \cdot V(X_i)\left[1 + (M-1)KR21\right]. \qquad (2)$$

Table 1. Item Sample and Total Test Statistics Obtained on the Otis-Lennon Test

Item Sample No.	Number of Pupils	Estimated Total Test Mean	Estimated Total Test Variance	
			Lord Estimate	Hill Estimate
1	136	19.41	18.36	17.59
2	137	17.26	27.32	25.27
3	127	16.61	36.92	33.86
4	126	15.24	21.29	13.99
5	127	16.06	18.43	16.00
Weighted mean over item samples		16.92	24.46	21.94

Total Test Results

Number of Pupils	Mean	Variance
602	17.08	29.05

Table 2. Item Sample and Total Test Statistics Obtained on the ERIE Science Process Test

Item Sample No.	Number of Pupils	Estimated Total Test Mean	Estimated Total Test Variance	
			Lord Estimate	Hill Estimate
1	88	14.24	48.88	44.09
2	93	17.76	38.64	24.68
3	89	10.78	11.21	6.98
4	101	13.35	50.09	46.52
5	92	14.46	28.13	24.06
6	91	14.62	17.65	10.24
7	99	11.67	13.41	-7.85
Weighted mean over item samples		13.98	29.72	21.25

Total Test Results

Number of Pupils	Mean	Variance
602	14.16	20.70

Table 3. Item Sample and Total Test Statistics Obtained on the Stanford Science Achievement Subtest.

Item Sample No.	Number of Pupils	Estimated Total Test Mean	Estimated Total Test Variance	
			Lord Estimate	Hill Estimate
1	65	42.31	146.01	136.74
2	69	39.13	208.24	200.98
3	63	33.65	96.11	81.28
4	61	30.82	154.25	141.03
5	72	37.22	149.59	136.56
6	70	41.14	74.61	57.72
7	62	44.19	69.11	55.25
8	60	39.33	107.68	89.16
9	67	31.04	89.64	81.71
10	64	38.28	111.70	103.14
Weighted mean over item samples		37.71	120.69	108.36

Total Test Results

Number of Pupils	Mean	Variance
602	36.44	108.16

F-ratios computed to test for statistically significant differences between the means of the testing conditions were .95, .30 and 4.40 for the three tests, respectively, with 1 and 1253 degrees of freedom. (More properly, the F-ratios should be calculated with class as the experimental unit, since subjects were not randomly assigned to classes. But since the intent of these calculations is to show the high degree of similarity between the two sampling results, the more conservative approach of using subjects as the experiment units is used). These results show no matrix sampling bias for the first two tests, and a very slight bias for the third test.

F-ratios computed to test for statistically significant differences between the obtained variance and the Lord estimate of the variance were 1.43, 1.19 and 1.12 for the three tests respectively, with 601 and 652 degrees of freedom. These results are statistically significant at the .05 level, two-tailed, for the first test only. F-ratios computed using the Hill estimate of the variance were 1.03, 1.32 and 1.00 for the three tests, respectively. These results are statistically significant for the second test.

Discussion and conclusion. The results indicate that matrix sampling can be practically applied when care is taken to minimize violation of assumptions. The very close matchups of means on the Science Process Test and the Otis-Lennon reflect this. The higher mean obtained from the matrix-sampled pupils on the Stanford may well emanate from a violation of the assumptions discussed earlier. The Stanford was given last in all cases, after two

tests the previous day. Observation indicated that both pupils and teachers were test-weary, and it was not unexpected to find lower mean scores on this test from the total test group. Had the administration of the Stanford been delayed for perhaps a week, the results may have correlated as well with matrix sampling estimates as did the other two tests.

The results concerning the variances were very much as expected. Previous studies had produced results which indicated that variances could be estimated well by matrix sampling. Both the Lord and Hill estimates were effective in estimating total-test variance in two of the three cases.